DOCUMENT RESUME

ED 421 526 TM 028 851

AUTHOR Parshall, Cynthia G.; Davey, Tim; Nering, Mike L.
TITLE Test Development Exposure Control for Adaptive Testing.
PUB DATE 1998-04-00
NOTE 30p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 12-16, 1998).
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; *Computer Assisted Testing; *Efficiency; *Selection; Simulation; Test Construction; *Test Items; Testing Problems
IDENTIFIERS *Item Exposure (Tests); Test Security

ABSTRACT

When items are selected during a computerized adaptive test (CAT) solely with regard to their measurement properties, it is commonly found that certain items are administered to nearly every examinee, and that a small number of the available items will account for a large proportion of the item administrations. This presents a clear security risk for testing programs that are available on more than a few scheduled testing dates throughout the year. Several approaches to this concern control item exposure rates through probabilistic mechanisms built into the selection process. While many of these exposure control procedures are quite effective in limiting rates of item use, they are also problematic to some extent. Several exposure control algorithms are described, including the rationale for their application and the nature of any inherent problems. An empirical comparison of the relative effectiveness of these methods is presented, based on simulated CATs. The unconditional Sympson-Hetter method (J. B. Sympson and R. Hetter, 1985) (USH), the conditional Sympson-Hetter method (M. Stocking and C. Lewis, 1995) (CSH), and the Davey-Parshall (T. Davey and C. Parshall, 1995) (DP) methods outperformed a no exposure control method, and the CSH and DP methods generally outperformed the USH method. The different targets and results of the CSH and DP methods are discussed. (Contains 1 table, 7 figures, and 17 references.) (SLD)

Test Development Exposure Control for Adaptive Testing

Cynthia G. Parshall
University of South Florida

Tim Davey
ACT, Inc.

Mike L. Nering
ACT, Inc.

Abstract

When items are selected during a computerized adaptive test (CAT) solely with regard to their measurement properties, it is commonly found that certain items are administered to nearly every examinee and that a small number of the available items will account for a large proportion of the item administrations. This presents a clear security risk for testing programs that are available on more than a few scheduled test dates throughout the year. Several approaches to this concern control item exposure rates through probabilistic mechanisms built into the item selection process. While many of these exposure control procedures are quite effective in limiting rates of item use, they are also problematic to some extent. We briefly discuss the need for exposure control, and then describe several exposure control algorithms, including the rationale for their application and the nature of any inherent problems. Finally, we provide an empirical comparison of the relative effectiveness of these methods.

Paper presented at the symposium *Adaptive Testing Research at ACT* at the annual meeting of the National Council on Measurement in Education, San Diego, 1998.

Test Development Exposure Control for Adaptive Testing

Introduction

Computerized adaptive tests are efficient because they successively select items that provide optimal measurement at each examinee's estimated level of ability. However, operational testing programs must usually consider additional factors in item selection. In practice, items are generally selected with regard to at least three, often conflicting goals: 1) to maximize test efficiency by measuring examinees as quickly and as accurately as possible, 2) to protect the security of the item pool by controlling the rates at which popular items can be administered, and 3) to assure that the test measures the same composite of multiple traits for each examinee by balancing the rates at which items with different content properties are administered.

This paper focuses on the goal of controlling item exposure rates, although the procedures detailed here would typically be used in conjunction with procedures for optimal item selection and balancing item content. A number of algorithms for controlling item exposure rates have been developed, and while many of them are quite effective in limiting rates of item use they are also problematic to some extent. We briefly discuss the need for exposure control, and then describe several exposure control algorithms, including the rationale for their application and the nature of any inherent problems. Finally, we provide an empirical comparison of the relative effectiveness of these methods.

Controlling Item Exposure Rates

When items are selected solely with regard to their measurement properties, it is commonly found that certain items are administered to nearly every examinee. Furthermore, a small number of the available items will account for a large proportion of the item administrations. This presents a clear security risk for testing programs that are available on more than a few scheduled test dates distributed throughout the year. The concern is that frequently administered items will quickly become compromised and no longer provide valid measurement. Some general approaches for addressing this concern include:

1. Use, and possibly even disclose, enormous item pools containing in excess of 5,000 items. Such pools could also be organized into subpools used on a revolving schedule so as to minimize the possibility of the same items reappearing in the same time period or geographical area.

2. Restrict testing to certain time "windows". This approach stops short of full testing on demand, but offers examinees greater flexibility in scheduling test dates than could be provided with most conventional tests.

3. Directly control item exposure rates through a statistical algorithm incorporated in the item selection procedures.

The use of either the "big pool" or restricted testing windows approaches will be largely dictated by practical and policy issues. However, in any case some means of directly controlling item exposure will likely be necessary. Neither large item pools or restricted testing windows are alone sufficient to ensure integrity of the item pool. Accordingly, a number of statistical procedures for controlling item exposure rates have been devised.

One simple method that has been recommended early on, is the so-called 4-3-2-1 procedure (McBride & Martin, 1983). This simple procedure has an item selection algorithm identify not only the best (most informative) item for administration at a given point, but the second, third, and fourth best items as well. Item exposure is then limited by allowing the best item to actually be administered only 40% of the times it is selected. The second, third and fourth best items are presented 30%, 20% and 10% of the times, respectively. This method is reasonably easy to implement, but provides limited protection against overexposure of those items that are more "popular", or likely to be selected for administration.

Another approach, the Sympson-Hetter method (Sympson & Hetter, 1985), was developed to provide more specific exposure control, through the use of *exposure parameters*. These exposure parameters are obtained through simulations conducted in advance of operational testing. Procedures have also been developed that build on the general Sympson-Hetter framework, but which are conditional on examinee ability (Stocking & Lewis, 1995b; Thomasson, 1995). In these conditional Sympson-Hetter approaches, a matrix of item exposure parameters is produced, with differing exposure parameters for each item, at each of a number of discrete ability levels. Finally, an alternative, conditional approach has also been investigated (Davey & Parshall, 1995). This approach does not condition on ability, but rather on the items which have already appeared during a given CAT. The rationale for this method is that item pool security may be best protected by directly limiting the extent that tests overlap across examinees.

Purpose

These procedures are all designed to provide statistical control to limit excessive item exposure; however, further investigation is needed to confirm their utility in practice. The purpose of the present

study was to investigate the effectiveness of several exposure control approaches in a realistic CAT setting. The Sympson-Hetter, Conditional Sympson-Hetter, and Davey-Parshall methods, as well as CAT administration under no exposure control and completely random item selection, were simulated and then compared in terms of their relative performance. Test length and precision were constrained; these methods were compared in terms of total pool usage, item exposure rates, and test overlap.

Details of Exposure Control Techniques

Specific details about each of the exposure control methods investigated in this study are provided below. The rationale for each approach, how any necessary simulations are conducted, and how the procedure is implemented operationally are addressed, along with any problems found with each method.

No Exposure Control. When no statistical exposure control is used, problems with uneven exposure may be anticipated, as indicated above. Some small proportion of the available items will be administered with excessive frequency, while others may not be administered at all. This may lead to serious security concerns, as well as wasting a fairly large proportion of the item pool. The use of no exposure control was used in this study to provide a baseline comparison.

Random Selection. The simple random selection of items for administration to examinees might be anticipated to provide far better exposure control than other methods. Of course, this is accomplished at the expense of measurement precision. Like the no exposure control method listed above, this method is provided in this study as a baseline for comparison, not as a recommended approach.

Sympson-Hetter. The Sympson-Hetter method for item exposure control (Sympson & Hetter, 1985), as well as the procedures addressed below, all make use of item *exposure parameters*. These must be obtained in advance of operational testing, through simulations. Each item is assigned a specific exposure parameter that has a value between zero and one. During operational exams, selected items are only administered with the probability given by their exposure parameter, as discussed more fully below.

This approach provides an improvement over the 4-3-2-1 method described above, in which any item selected as "best" at a given point in the CAT is automatically assigned an exposure probability of 40%. The unique exposure parameter assigned specifically to each item in this method is designed to provide better results, protecting an item more or less from overexposure, based on the frequency with which that item tends to be selected for administration.

The exposure parameters for the individual items are obtained through simulations, conducted prior to operational testing. First, a "maximum exposure rate" is set. For example, it may have been decided

that no item should be administered to more than 15% of all examinees. This target maximum rate is somewhat subjective and is a function of pool size, average test length, and desired level of security.

Several thousand adaptive test administrations are then simulated. For the first simulation, exposure parameters are set at 1, and every item which is selected is also administered. Following each simulation, the number of times each item was selected and the number of times it was actually administered are tallied. The number of times an item is administered is then compared to the target maximum exposure rate. The exposure parameters for those items whose frequency of administration exceeds the target maximum exposure rate are successively adjusted downward. For example, this may be accomplished by multiplying the current value by .95. An item that appears less frequently than permitted may be adjusted upward (to a maximum of 1), by multiplying the current value by 1.04. This upward adjustment keeps an exposure parameter from being overly reduced due to its chance over-frequent occurrence in a single simulation. These adjustments continue through the cycle of thousands of simulations. The cycle ends when the exposure parameters have stabilized and no items exceed the target maximum exposure standard.

The final result of this series of simulations is an n-long vector (where n is the number of items) of exposure parameters. These parameters are used in operational testing by incorporating them into the item selection algorithm. That is, during the actual CAT, an item is identified as the best item to be administered next to a given examinee. This may be done according to maximum information, minimum posterior variance or any other primary or optimal item selection criterion; content constraints may also be factored into the item selection process. A uniform (0,1) random number is then generated, and compared to the selected item's exposure parameter. If the exposure parameter for the selected item is greater than the random number, the item is administered. Otherwise, the item is not administered, and a new item must be selected.

An item is thus typically only administered a given proportion of the times when it is selected, with that proportion specified by the item's exposure parameter. Those items that tend to be frequently selected are assigned more restrictive exposure parameter, while items that are less frequently selected have more relaxed exposure parameters. An extreme case might be an item with an exposure parameter such as .98, which would allow the item to be administered almost every time it was selected. Items selected but not administered are set aside and rendered unavailable for reselection, unless the item pool becomes exhausted.

Conditional Sympson-Hetter. Extensions to the original Sympson-Hetter approach have been developed (Stocking & Lewis, 1995b; Thomasson, 1995) to correct for a problem the method often has in practice. This problem is that the exposure parameters obtained in the Sympson-Hetter method are dependent upon the expected distribution of examinee ability used in the simulation phase. This distributional dependency is addressed in the Conditional Sympson-Hetter through the use of different item exposure parameters for different ability levels.

The unconditional Sympson-Hetter (USH) uses a single, global exposure parameter for each item. This unconditional exposure parameter is obtained during the initial simulation phase; thousand of CATs are simulated, based on given conditions such as test length, reliability, etc. The simulations must also use an expected distribution of examinee ability. However, the actual observed distribution may differ from this expected examinee distribution, either for the total operational group or for some sub-group of examinees. To the extent that the simulated examinee distribution is a poor match to the operational examinee distribution, exposure control will not be maintained as desired.

Items will be selected in actual operation at frequencies that differ from those expected; thus, the exposure parameters for individual items will be set at inappropriate rates. Items that are selected operationally more frequently than expected may have overly relaxed exposure parameters, resulting in actual item exposure that is greater than the target maximum. Items that are selected operationally less frequently than expected may have overly restrictive exposure parameters, providing protection where it is unneeded and limiting test efficiency.

A related problem with the USH approach results from its tendency to produce very relaxed exposure parameters for the items that are informative in the distribution tails (or wherever few people are expected to be). This can be a problem, because while those items may be selected for few people overall, they will tend to be selected for almost all the examinees who have that level of ability. An unconditional approach to exposure control will both select and administer many of those items to that set of examinees. This results in very poor exposure protection, for example, in test-retest situations. Additionally, there is some reason to expect that examinees will have friends of similar ability, and that if they cheat, or share items, it is with these similar-ability test-takers. Unconditional exposure parameters would put the security of the item pool at risk in both of these scenarios.

For all these practical reasons, a conditional approach to exposure control may be desirable. The conditional Sympson-Hetter (CSH) is designed to directly control item exposure to examinees with

similar abilities, and to provide exposure parameters that are independent of any particular ability distribution.

Exposure parameters are estimated for this method by a cycle of simulations. In this initial simulation phase, the ability range is divided into a number of discrete classifications (m). The frequencies of item selection and administration are kept separately for each of the m levels. Exposure parameters are computed for each of these discrete levels. While the USH procedure yields a vector of exposure parameters, the CSH approach produces a matrix of exposure parameters. This is an $n \times m$ matrix, where n is the number of items and m is the number of discrete ability levels; within each level of theta, each item has a specific, conditional exposure parameter.

In operational testing, the examinee's current estimate of theta is used to determine which m column of the matrix should be used to obtain an item exposure parameter. The conditional exposure parameter then functions like an USH exposure parameter. It is compared to a uniform random number, and limits actual administration of the selected item to the specified rate.

The CSH method is designed to overcome some important, practical limitations of the USH method. It provides exposure parameters that are independent of the expected examinee distribution, and it allows direct control over exposure for different ability levels. Additionally, Stocking and Lewis (1995) point out that, if desired, different target maximum rates can be set for the different m levels. However, the method is more complex than those described previously. And, it is more time-consuming to conduct the initial simulations. Finally, the method has not been fully tested on a range of item pools, with varying restrictions and characteristics.

Davey-Parshall. An alternative, conditional approach has also been investigated (Davey & Parshall, 1995). This approach, however, is not conditional on ability, but rather on the items which have already appeared during a given CAT. The Davey-Parshall method addresses exposure control by limiting the extent that tests overlap across examinees. The goal of this exposure control procedure is not to strictly limit the frequency of item use, but rather to directly minimize the extent that tests overlap across examinees or testing occasions.

Given that adaptive tests without exposure control will often result in clusters or sets of items which appear together with unwelcome frequency, the Davey-Parshall (DP) method provides exposure probabilities conditioned on those items that have already appeared. In any testing application, individual items can become compromised through overuse. However, the adaptive testing process is fairly robust to the effects of a few isolated spuriously correct responses (Davey & Miller, 1992). The

danger that the DP approach addresses is, instead, tests that overlap to the point where examinees can disclose large proportions of their tests to one another, or situations in which an examinee being retested is presented with substantially the same items as on the previous test. Rather than focusing on examinee ability, the DP method is designed to control item exposure by directly controlling test overlap. Thus, if two or more items tend to appear together exceptionally often unconditionally, the DP procedure will limit the probability of the remaining items in this set appearing once any item in the group has been administered.

The DP method, like both the USH and the CSH, utilizes exposure parameters that are obtained in advance of operational testing through a series of simulations. As is true for the USH and CSH methods, a maximum desired exposure rate for individual items is initially determined. Unlike the other two procedures, the DP method also uses a maximum exposure rate for pairs of items. This is an upper limit on the frequency with which pairs of items ought to co-occur.

The DP procedure utilizes an $n \times n$ matrix of exposure parameters, where n is the number of items in the pool. Diagonal elements of this table contain unconditional exposure parameters for individual items, similar to those used in the USH method. The off diagonal entries are the conditional parameters that control the frequency with which item pairs or clusters can co-occur. Sympson-Hetter can be viewed as a special case of the proposed procedure, in which all of the off-diagonal values are set to unity.

As is conducted for the USH and CSH procedures, a series of simulations is carried out to determine the values of the individual exposure parameters. In the first simulation, all of the exposure parameters (both individual and pairwise) are set to one. Following each simulation, the number of times each item and pair of items appears is tallied. The individual exposure parameter for any item that appears more frequently than the target maximum is adjusted by multiplying the current value by .95; an item that occurs less frequently than is allowed is adjusted upward (to a maximum of 1) by multiplying by 1.04.

Adjustment of the pairwise exposure parameters is a little more complex. First, the pairwise item frequencies are summarized in contingency tables like the one below. A statistical test is then used to determine whether the co-occurrence of a pair of items is greater than would occur by chance. If the items display a pairwise association that is greater than desired, the pairwise exposure parameter for that item pair is adjusted downward.

		Item I		
		Appeared	Did not Appear	Total
Item j	Appeared	A	B	A+B
	Did not Appear	C	D	C+D
	Total	A+C	B+D	A+B+C+D

Cell A for a given contingency table reports the number of tests in which both items in the pair appeared. Ideally, this pairwise frequency should be no larger than would be expected by chance, indicating no positive association between the pair of items. The DP method tests this hypothesis of no positive association using a modified chi-square statistic. The chi-square test is modified to ignore negative associations, which are in some sense desirable. The statistic used was:

$$\chi^2 = \begin{cases} \dfrac{(A+B+C+D)\bullet(A\bullet D - B\bullet D)^2}{(A+C)(B+D)(C+D)(A+B)} & if\ (A\bullet D - B\bullet D) \geq 0 \\ 0 & if\ (A\bullet D - B\bullet D) < 0 \end{cases}$$

If the statistic for a given item pair exceeded a specified maximum value, the i,j entry of the exposure table for that pair of items may be decreased by being multiplied by .95. As with the individual exposure parameters in the table diagonal, entries may be increased (multiplied by 1.04) for those item pairs whose statistic fall below the maximum value.

Clearly, the problem of multiple comparisons must be considered when specifying a maximum value for the chi-square statistics. Some experimentation with the item pool used in this study suggested a value of 8.28 yielded desired test conditions. However, additional experimentation would be required before values appropriate for other item pools, other test lengths, other ability distributions, and other security goals could be determined.

The simulations continue, with adjustments to both individual and pairwise exposure control parameters, until the matrix stabilizes at the desired exposure rate. The final matrix can then be used in operational testing. Operational use of the conditional procedure can be illustrated using the following small example of an exposure table.

		Item				
Item		**1**	**2**	**3**	**4**	**5**
	1	.8	.8	.2	1.	.4
	2	.8	1.	1.	1.	1.
	3	.2	1.	.4	.8	.6
	4	1.	1.	.8	1.	1.
	5	.4	1.	.6	1.	.7

The exposure table is symmetric, with values above the diagonal equal to those below. As described above, the diagonal elements are the individual exposure parameters. Item 3, for example, has a diagonal entry of .4, suggesting that it is individually selected overly often and should be administered much less often than it is selected. The off-diagonal entries are the pairwise exposure parameters; low off-diagonal values are used to prevent popular pairs or sets of items from co-occurring. The item 1-3 combination is seen to be especially troublesome and will therefore be strictly controlled. However, item 4 is apparently not closely associated with other items in the table and so can freely appear with any.

During operational testing the exposure table would be used to determine conditional exposure parameters through the following procedure. First, an item would be selected for administration based on any appropriate item selection procedure. Next, this item's entries in the exposure table are obtained. The individual exposure parameter, from the diagonal, is denoted by e_{ii}. The pairwise exposure parameters, from any off-diagonal entries involving both the selected item and previously administered items, are denoted by e_{ij} (where j ranges from one to the number of items so far administered). Finally, the conditional probability of administering the selected item is computed by taking the mean of the set of e_{ij} values and multiplying the result by e_{ii}. The selected item is administered with this probability.

These individual and pairwise parameters are used in concert to determine the actual administration of any given item. This is dependent upon the selected item's individual frequency of selection, and it's tendency to be selected in conjunction with any specific items that have already been administered to the

examinee. For example, in the sample table, suppose that item 5 has been selected and that items 2 and 3 have already been administered. Then $e_{55} = .7$, and $e_{52} = 1.$, $e_{53} = 6$. The conditional probability of administering item 5 given that items 2 and 3 have already appeared is thus: .7*(1 + .6)/2 = .56. Item 5 would therefore be administered just over half the times it was selected, if items 2 and 3 had already been presented to the examinee.

The DP method, like the CSH approach, is intended to overcome weaknesses found in earlier exposure control methods. It is designed to lead to a more balanced use of the item pool and to reduce the overlap in tests both for repeat test-takers (or examinees with similar ability) and across peers (or examinees of differing ability). However, this potential improved performance comes at the cost of greater complexity and a more time-consuming simulation phase. Further evidence of its effectiveness is also needed.

Methods

Data Generation

The exposure control procedures detailed above were investigated in this study through simulated CATs. Simulated item responses were generated from a multidimensional item response theory (MIRT) model. This model included not only the major dimensions that provide basic structure, but also numerous minor dimensions that are characteristic of actual data. MIRT data generation provides simulated data that is more similar to real data than that produced by more typical unidimensional IRT models (Davey, Nering, & Thompson, 1997; Parshall, Kromery, Chason, & Yi, 1997).

Multidimensional simulation begins by fitting a MIRT model to a sample of actual data. In this case, the data consisted of eight randomly equivalent groups of 3,500 examinees who responded to each of eight parallel forms of a large scale assessment. The eight datasets were independently calibrated in 50 dimensions using a modified version of NOHARM (Fraser, 1986; Fraser & McDonald, 1988). No attempt was made to interpret the resulting solution; rather, the fitted model was treated only as a template for generating data. Even though the examinee samples were randomly equivalent it was still necessary to rotate the resulting NOHARM parameter estimates to a common ability metric [see Thompson, Nering, & Davey (1997) for details of this procedure]. Two additional test forms were cloned to form a total of 10 test forms comprising 600 total items.

Examinees were simulated by drawing ability vectors from specified population distributions. For some analyses, simulated examinees were drawn from a multivariate normal distribution (N50(0,I)). For other analyses, it was desirable to have a set of examinees that produced a roughly uniform

distribution of number right scores. Ability vectors for these analyses were drawn from a series of mixing distributions. [For a description of this second sampling method see Nering, Thompson, and Davey, (1997)].

The set of MIRT item parameters and examinee abilities was then used to generate data, both for determining exposure control parameters and for administering operational tests. Item responses were generated by determining the probability of a correct response on a given item, for a given examinee, and then comparing that probability to a random number sampled from a uniform (0,1) distribution. If the random number was less than or equal to the probability of a correct response, then the response was scored correct; otherwise, the response was scored incorrect.

Features of the CATs

The simulated CATs were of variable lengths, with a minimum of 15 items and a maximum of 45 items. By study design, the parameters of each exposure control procedure (e.g., target maximum exposure) were set so as to produce comparable conditional standard errors of measurement and average test lengths of approximately 30 items. CAT designs may vary in the three, interacting features of test length, reliability, and item exposure. In this study, we controlled test length and reliability, where possible, in order to make direct and valid comparisons of item exposure across exposure control methods.

Each test ended when a specified amount of information had been accumulated at the examinee's estimated ability level. The information target was set to cause the CAT to approximate the measurement precision of the conventional forms that comprised the item pool. Items were selected by maximum information, with selection constrained by content considerations (Davey & Thomas, 1996). Provisional ability estimates were computed by Owen's bayes mode approximation (1969, 1975), while final estimates were obtained using maximum likelihood (MLE).

Conditions of Study

The five exposure control conditions investigated in this study included two baseline conditions, no exposure control (NOEXP) and random item selection (RAND). The three more promising exposure control procedures were an unconditional Sympson-Hetter procedure (USH), a conditional Sympson-Hetter approach (CSH), and the Davey-Parshall (DP) method. These three methods were constrained so that that their average test lengths and conditional standard errors of measurement (CSEM) were as similar as possible, in order to make appropriate comparisons of their effectiveness at controlling item exposure. However, it is not possible to constrain the remaining two method to both of these goals

concurrently. Either test length or test precision can be targeted in the NOEXP and RAND methods, but not both. For this study, the NOEXP condition was set to maximize precision (or, to minimize standard error of measurement); test length and, of course, exposure control were not constrained. The RAND method, on the other hand, was targeted to administer exactly 30 items in a test, randomly drawing items from the pool. No other constraints were used; test precision, maximum information, and content balancing were all disregarded.

Item usage rates for the total pool were examined by administering tests to 100,000 examinees sampled from a multivariate normal distribution. This sampling procedure resulted in a dataset very much like that which would be observed in an operational testing situation.

A second set of simulations sampled multidimensional abilities so as to produce roughly equal numbers of examinees at each number right score level. These data from uniform distribution of number right scores were used when results were analyzed conditionally on score level. That is, the uniformly distributed data were used to compute test length, standard errors of measurement, and test overlap rates. All of these analyses were examined conditional on raw score levels.

Test lengths and standard errors of measurement were based on 2500 examinees at each raw score level (with 50 score levels, this resulted in a total of 125,000 simulated examinees for each condition). Test overlap rates were based on samples of 200 examinees at each of the 50 score level. While this sample size (of 10,000) seems comparatively scant, overlap rates were determined by comparing each examinee's test with the tests of all other examinees. The total number of comparisons made is then seen to be enormous.

Results

Pool Usage/Exposure Rates

Total pool usage was examined by examining the actual exposure rate for each of the 600 items in the pool. These exposure rates were computed as the simple proportion of tests on which an item was administered. Table 1 shows the frequency of items by exposure rate, across the five studied conditions. The goal of good exposure control is use as much of the item pool as possible, without overly using any part of it.

The NOEXP condition, as was expected, shows the poorest performance. Over 500 items were administered on only 1% of the tests, or not used at all. A small subset (53) of the items that were used, however, were over exposed, with exposure rates of 10% or greater. This is the typical pattern found

under adaptive testing without exposure control. A minority of maximally informative items in the center of the distribution are used excessively, while a majority of the items remain unused.

The RAND method, on the other hand, shows a best case scenario for item exposure. The items are used with similar levels of frequency; all of them are used, but none is over-exposed. Exposure rates for items in this condition vary from a low of 2% to a high of 7% of the tests administered, with the majority of the items having exposure rates close to 5%. This excellent exposure control comes at a price, however. The random item selection procedure ignores critical CAT characteristics, including information, efficiency, and content.

The DP and USH methods perform similarly, emphasizing the relationship between them (i.e., that the USH can be seen as a special case of the DP in which all of the pairwise exposure parameters are set to 1). As expected, both of these methods outperform the NOEXP method. They both have far fewer unused items than the NOEXP method, without having any items in the highest exposure rate category. The DP method shows definite improvement over the USH, with fewer unused items, and more items in the moderate exposure control categories.

The CSH procedure provides exposure control for the pool somewhat differently. In this table, the CSH shows the best total pool usage in the low number of unused and almost unused items. More items have moderate exposure rates (2% to 7%) for this method than for any of the other four. However, the CSH has the highest number of items in the worst exposure rate category, of 10% and above. This may be the result of the fact that this method does not use any global exposure control, but rather controls exposure exclusively at specific ability levels.

Test Characteristics

Test length, conditional on number correct score, is reported in Figure 1 for the five exposure control procedures. Figure 2 shows the conditional standard errors of measurement (CSEM) for the studied exposure control procedures.

As discussed above, the three methods of greatest interest, the USH, CSH, and DP methods, were designed to have equivalent average test lengths and extremely comparable CSEMs. However, for the two baseline procedures, the NOEXP and RAND, a choice had to be made to control either test length, or precision, but not both. The relationship between test length, exposure rates, and standard error can be seen clearly in these two procedures. The RAND procedure was fixed at a test length of 30 items, while precision (and every other CAT features) was allowed to vary. This condition can be seen to have a constant test length (Figure 1), excellent exposure control (Table 1), and the poorest CSEM of the five

methods (Figure 2). The NOEXP procedure, on the other hand, used an item selection procedure that targeted maximum information and ignored all else. This resulted in a highly efficient test (Figure 1), with a moderate CSEM (Figure 2), and very poor exposure control (Table 1).

A comparison of the three remaining exposure control methods, the USH, CSH, and DP, shows highly similar patterns of conditional standard error. While these three methods all had average test lengths of 30 items, the pattern of test length across score level is not as consistent. The CSH method administered longer tests for examinees in the tails of the score distribution, and shorter tests for examinees in the middle of the distribution, as compared to the USH and DP methods. This conditional test length effect is a result of the CSH's explicit control of item exposure conditional on ability.

Test Overlap Rates

Finally, test overlap rates were determined for each pair of simulees using the uniform sampling with 200 simulees at each level. Test overlap is essentially the proportion of items that two examinees have in common. Obviously, the lower this number the better. Because this resulted in 200 x 200 comparisons at each score level, it was necessary to summarize these statistics. Thus, the median, 90th and 10th percentiles of the overlap distributions were plotted for each score level. This resulted in three overlap plots for each exposure control condition studied.

Figures 3 - 7 provide these plots for each condition. Beside each plot are three numerical summaries of the plot. The Overall value provides the percent of test items that any random two examinees might be expected to have in common. The Similar value provides the percent of test overlap for pairs of examinees whose total true scores were no more than two points apart, while the Different value reflects the average percent of test overlap for pairs of examinees whose scores were more than two points apart. (The Similar examinees fall in the diagonal of the plot, from the 0,0 point on the grid to the back corner of the plot; a tendency for greater test overlap in this group causes the "saddle-shape" of most of these plots.) The Similar value for the 90th percentile distribution represents the worst case scenario for test overlap, while the Overall value for the median distributions represents the average expected overlap rate between any two randomly selected individuals. The plots themselves are "unweighted"; the numerical summaries just described, however, are weighted by expected frequency of examinees at the various score points. Thus, the numerical summaries reflect a greater import for the effect of test overlap in the middle of the distribution, where more examinees are expected to fall, and less import for overlap in the distribution tails.

Figure 3 displays results for the NOEXP condition. As might be expected, this condition demonstrates very poor test overlap performance. The median distribution plot indicated that any two random examinees could be expected to have 30% of their test items in common. Examinees with similar abilities might have 64% items in common. Recall that the NOEXP condition had an average test length of close to 15 items; test overlap might be expected to be even worse for longer tests.

The RAND condition, displayed in Figure 4, shows a very different pattern of results. The test overlap results for this atypical condition show practically identical overlap rates for examinees, regardless of a given examinee's level of ability or a pair of examinees similarity of test score.

Once again, the USH and DP methods show a similar pattern of results, while the CSH differs. A comparison of the unweighted plots for the USH and DP methods reveals improved performance for the DP method over the USH, particularly in the tails of the score distributions. For example, for the USH median distribution, a pair of examinees at the high end of the score scale might be expected to have 30% test overlap (using the percentiles on the left side of the figure). For the DP method, that same examinee pair might have less than 25% test overlap. The numeric values for those same conditions (provided to the right of the figures) were weighted by expected frequency distributions, and emphasize examinees in the middle of the score distributions. Based on these figures, the USH outperforms the DP.

The CSH plots indicate that it provides greater control in the distribution tails, as compared to these other two methods. Both a visual examination of the plots and the similarity of the CSH values across the Overall, Similar, and Different cases suggest that the conditional exposure control exercised by this method results in test overlap rates which are conditionally consistent across test score. The weighted, numeric values show results that are very similar to the USH, and somewhat better than the DP.

Conclusion

Protecting the integrity of the item pool is a critical issue in computer adaptive testing. This will become even more crucial as increasing numbers of high stakes testing programs move to the CAT environment. While procedures for successfully controlling item exposure rates have been developed, none has been demonstrated to have generally superior performance.

The USH, CSH, and DP methods all outperformed the NOEXP control condition in terms of item protection. Indeed, the NOEXP condition is so poor, under utilizing 85% of the item pool, as to be unacceptable for most testing applications. The RAND method, of course, was used only as another

baseline. Despite its excellent exposure control, it is also unacceptable, due to its exclusion of other essential item selection criteria.

Both of the conditional methods examined here, the CSH and the DP, generally outperformed the older, unconditional method (the USH). However, the exposure protection they provide targets different goals, and yields different results from one another. In terms of pool usage, the CSH resulted in lower average item exposure rates than the DP, but also had more excessively used items, when results are considered globally. A consideration of test overlap rates showed the CSH to provide generally better performance than the DP in the distribution tails; however, this improvement over the DP is not as clearly evident in the middle of the distribution.

The better test overlap protection provided by the CSH in distribution tails came at the price of longer tests. Although this study controlled for average test length across all examinees, an examination of conditional test length revealed that the CSH administered longer tests to examinees in the tails of the score distribution. The DP method provided slightly poorer test overlap in the distribution tails, but accomplished this with shorter tests.

Certainly, an important limitation of this study is the use of only one item pool. Pool size and quality, as well as the requirements of a given CAT no doubt interact with the performance of exposure control methods. Characteristics of this pool and this CAT could have impacted the variations in performance for the CSH and DP methods. Further comparisons of these two promising methods under a wider variety of conditions is recommended. Finally, a hybrid procedure, that provides direct control over test overlap conditional upon ability is also worth investigation.

References

Brown, J.M. & Weiss, D.J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Psychometric Methods Program.

Davey, T. & Miller, T. R. (1992). *Effects of item bias and item disclosure on adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Columbus, OH.

Davey, T., Nering, M. L., & Thompson, T. (1997, June). *Realistic simulation of item response data.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Davey, T., & Thomas, L. (1996, April). *Constructing adaptive tests to parallel conventional programs.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Fraser, C. (1986). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* [computer program]. Center for Behavioral Studies, The University of New England, Armidale, New South Wales, Australia.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267-269.

McBride, J.R. & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.), *New horizons in testing* (pp.223-226). New York, Academic Press.

Nering, M., Thompon, T., & Davey, T. (1997, June). *Simulation of realistic ability vectors.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Parshall, C. G., Kromrey, J. D., Chason, W. M., & Yi, Q. (1997, June). *Small samples and modified models: An investigation of IRT parameter recovery.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Stocking, M. L., & Lewis, C. (1995a). *A new method of controlling item exposure in computerized adaptive testing.* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995b). *Controlling item exposure conditional on ability in computerized adaptive testing.* (Research Report 95-24). Princeton, NJ: Educational Testing Service.

Sympson, J.B. & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thomasson, G. L. (1995). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis.

Thompson, T., Nering, M., & Davey, T. (1997, June). *Multidimensional IRT scale linking without common items or common examinees.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Table 1

Total Pool Usage – Frequency of Items with Various Exposure Rates, Across Exposure Control Method

	Exposure Control Methods				
Exposure Rates	**NOEXP**	**RAND**	**USH**	**CSH**	**DP**
.10 +	53	0	0	111	0
.09 - .10	3	0	5	24	10
.08 - .09	1	0	210	32	188
.07 - .08	3	0	122	31	136
.06 - .07	2	105	5	29	11
.05 - .06	5	177	7	36	7
.04 - .05	7	231	10	43	11
.03 - .04	2	87	9	33	11
.02 - .03	5	0	11	44	14
.01 - .02	9	0	20	60	29
0 - .01	510	0	201	157	183

Note: The pool consisted of 600 items.

Figure 1

Average Items Administered

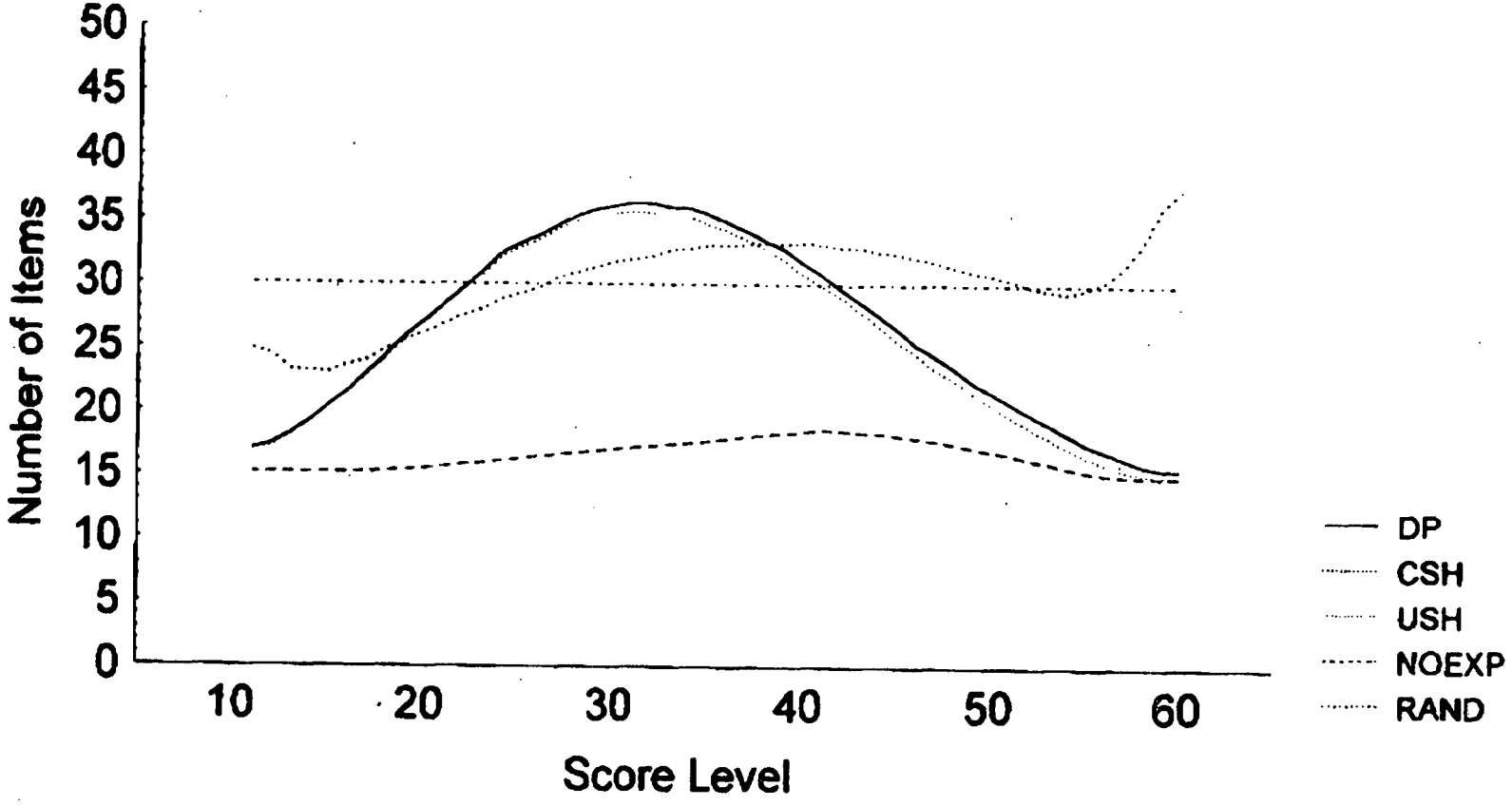


Figure 2
Conditional Standard Error

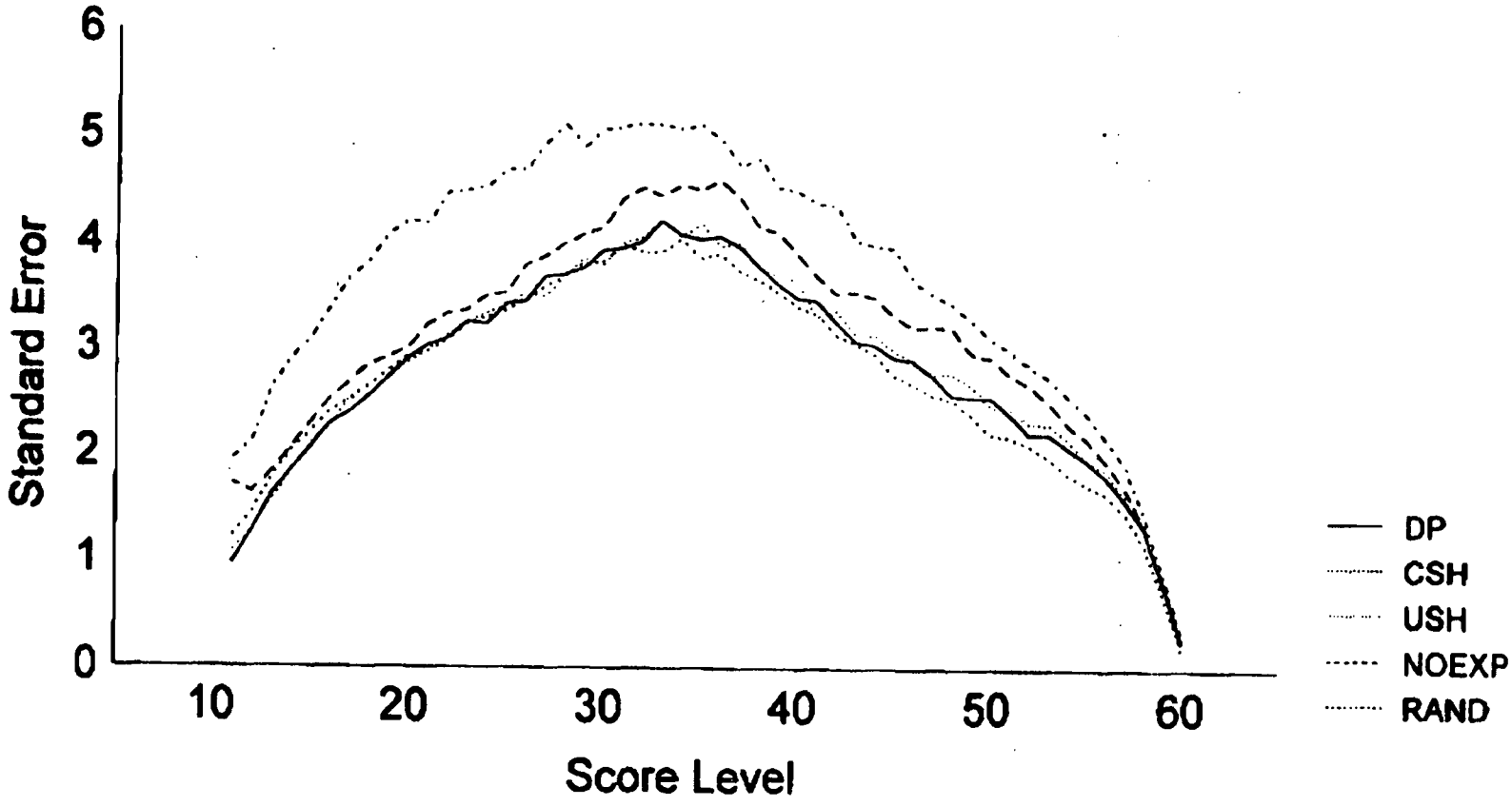
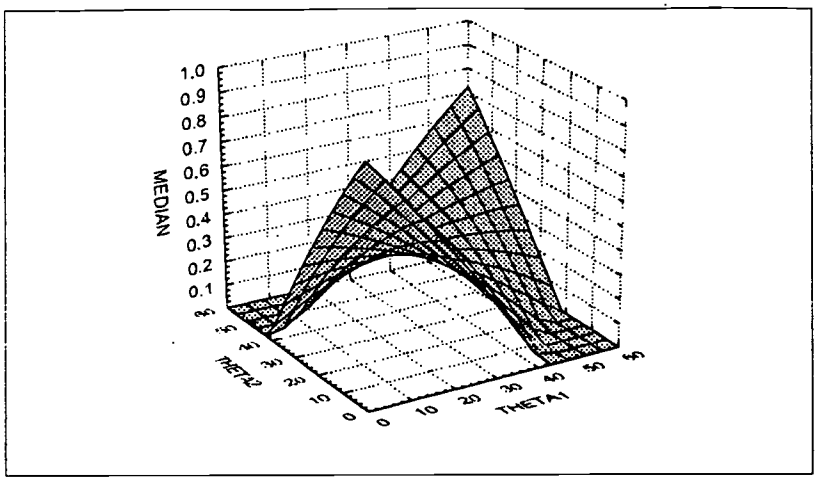


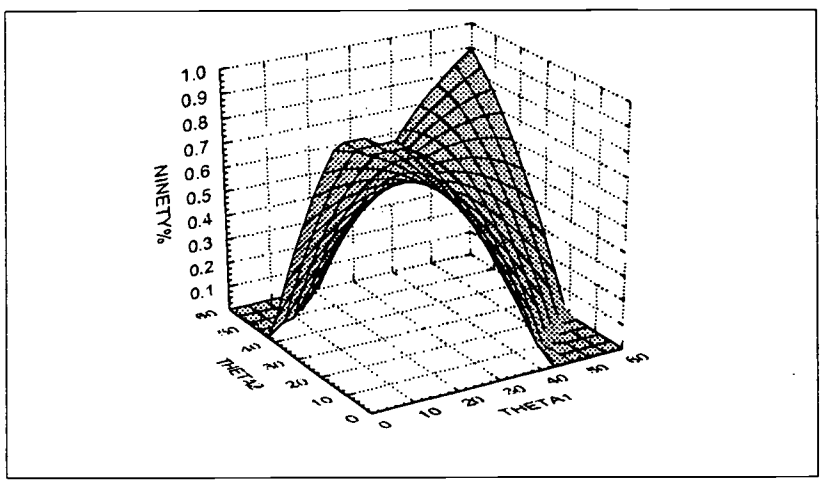
Figure 3

Overlap Rates with No Exposure Control



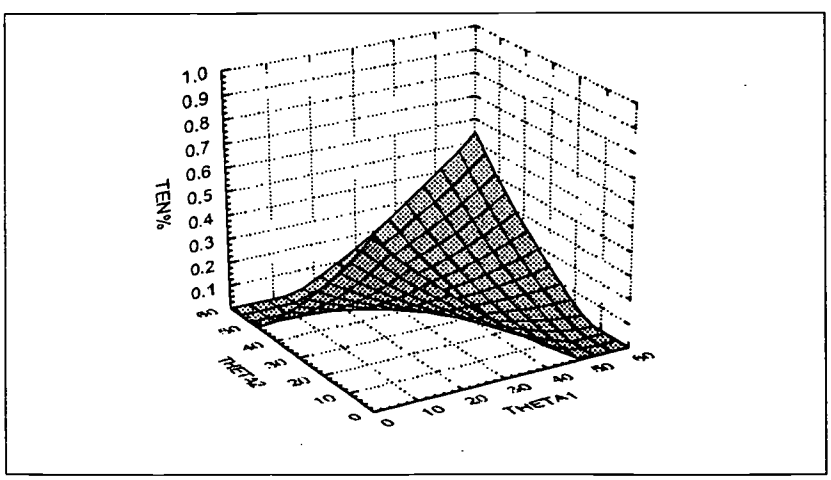
Median of Overlap Dist

Overall:	0.3057
Similar:	0.6410
Different:	0.2590



90th % of Overlap Dist

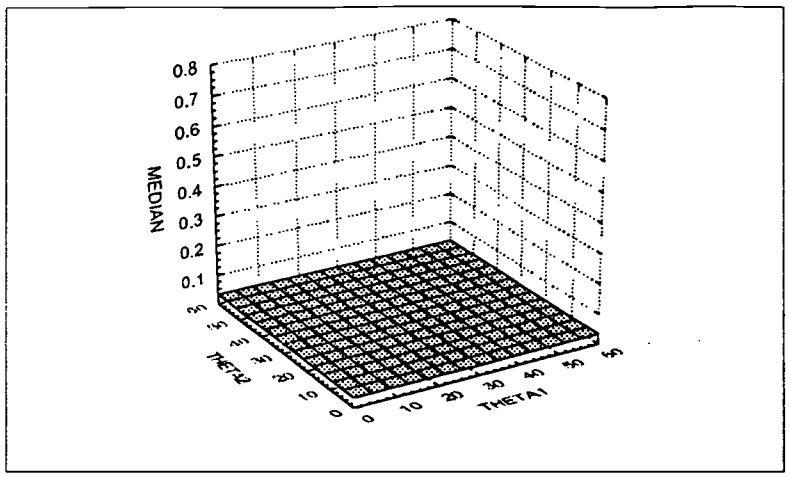
Overall:	0.5542
Similar:	0.8650
Different:	0.5110



10th % of Overlap Dist

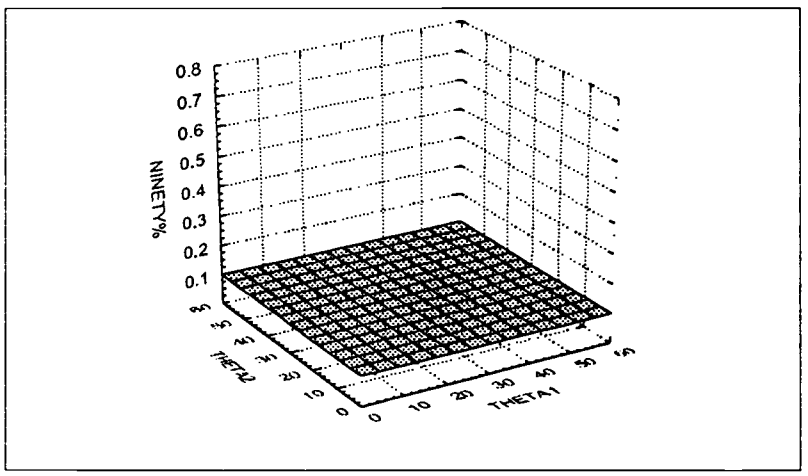
Overall:	0.1275
Similar:	0.2872
Different:	0.1053

Figure 4
Overlap Rates with Random Item Selection



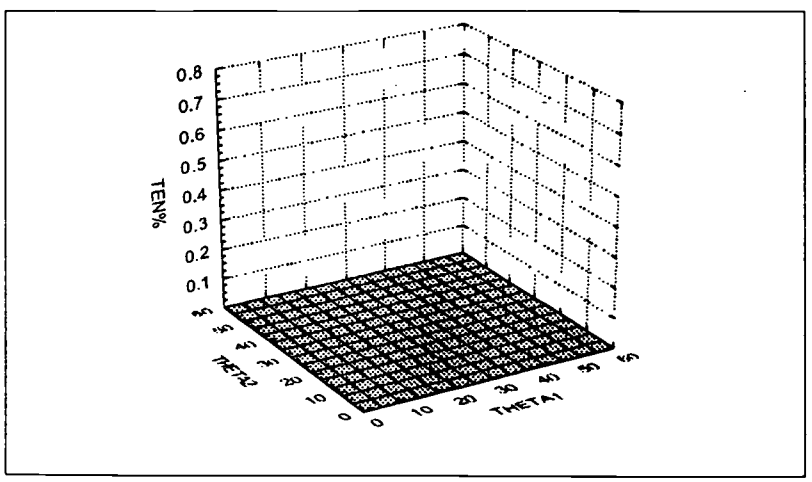
Median of Overlap Dist

Overall:	0.0330
Similar:	0.0333
Different:	0.0329



90th % of Overlap Dist

Overall:	0.0989
Similar:	0.0998
Different:	0.0988

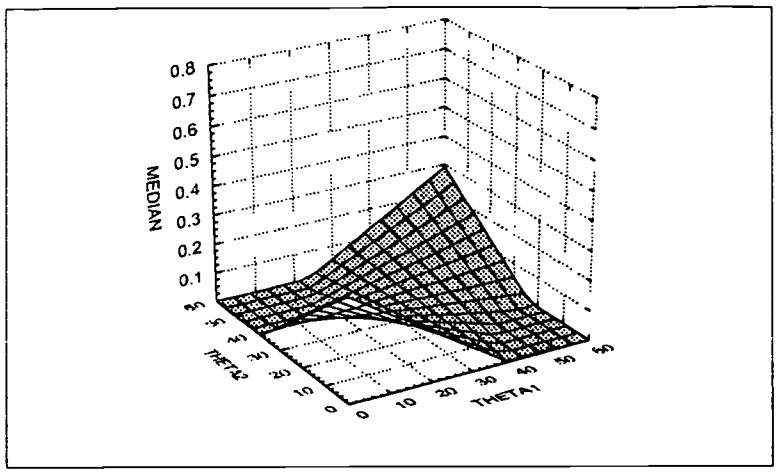


10th % of Overlap Dist

Overall:	0.0000
Similar:	0.0000
Different:	0.0000

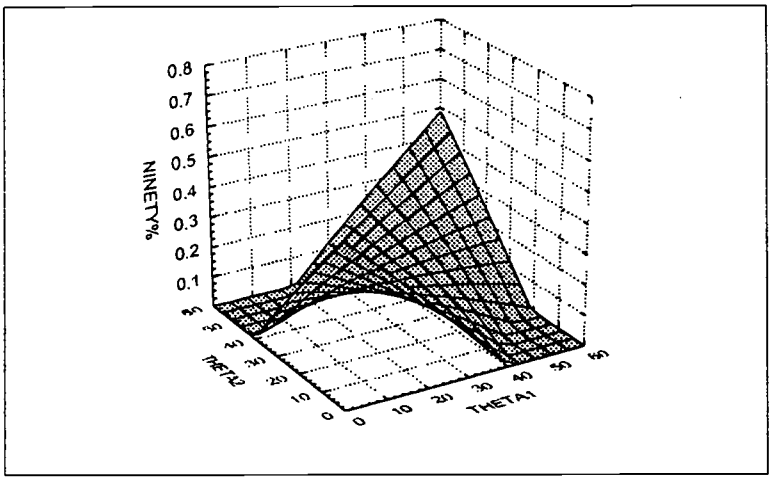
Figure 5

Overlap Rates for Unconditional Sympson-Hetter Exposure Control



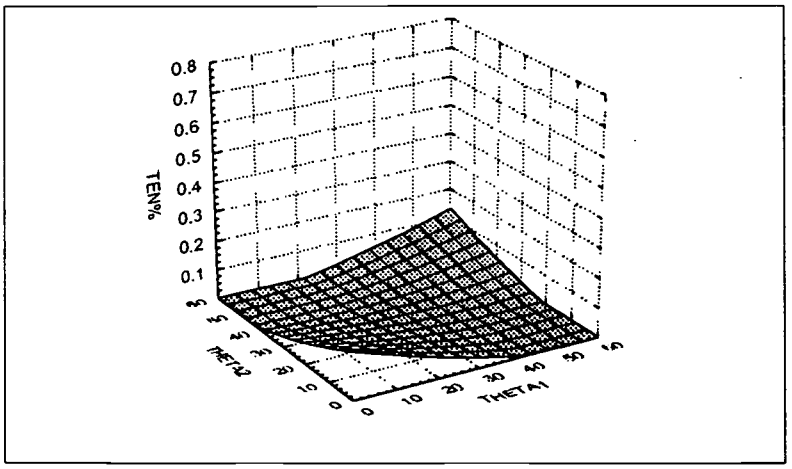
Median of Overlap Dist

Overall:	0.0656
Similar:	0.1359
Different:	0.0558



90th % of Overlap Dist

Overall:	0.1457
Similar:	0.2336
Different:	0.1335

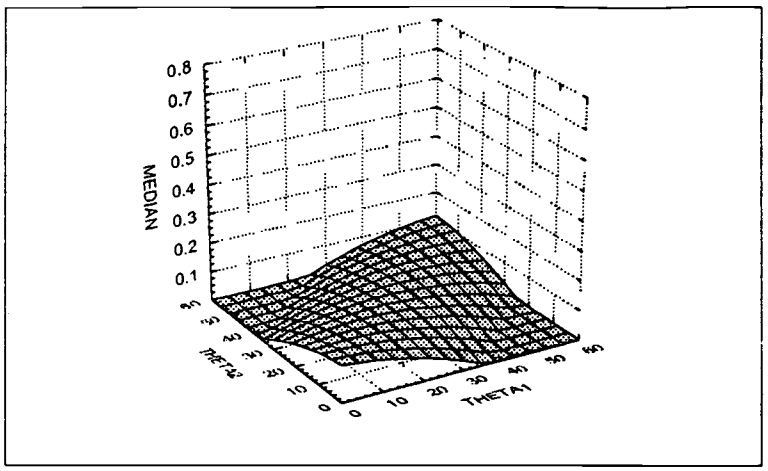


10th % of Overlap Dist

Overall:	0.0158
Similar:	0.0496
Different:	0.0111

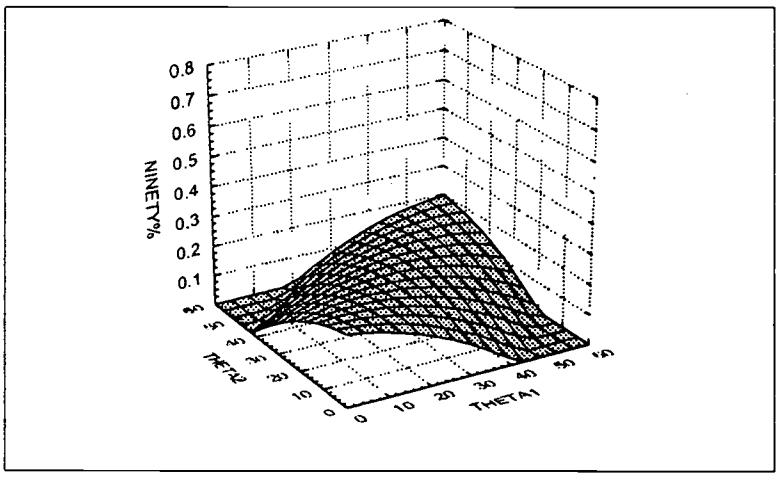
Figure 6

Overlap Rates for Conditional Sympson-Hetter Exposure Control



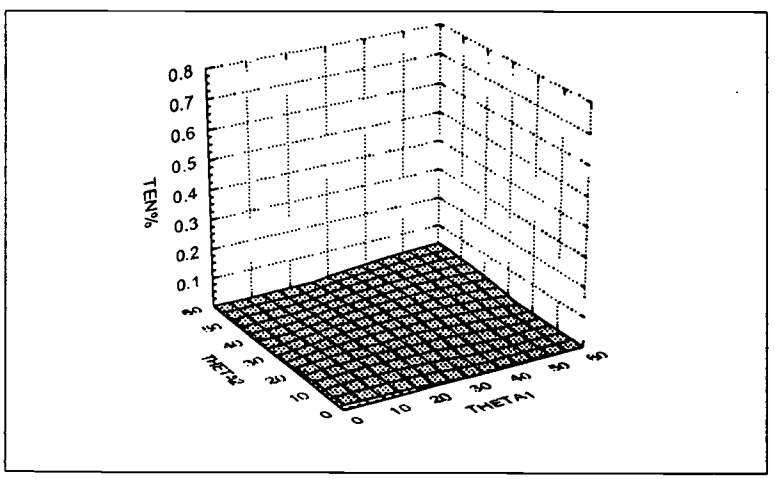
Median of Overlap Dist

Overall:	0.0683
Similar:	0.1214
Different:	0.0617



90th % of Overlap Dist

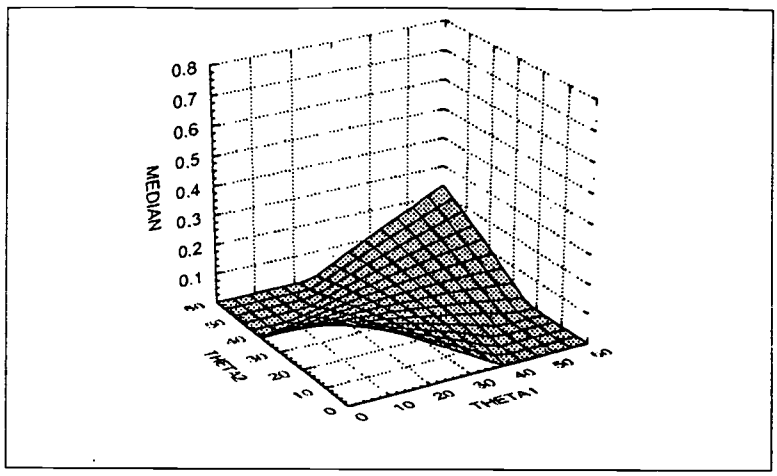
Overall:	0.1401
Similar:	0.2104
Different:	0.1312



10th % of Overlap Dist

Overall:	0.0195
Similar:	0.0486
Different:	0.0159

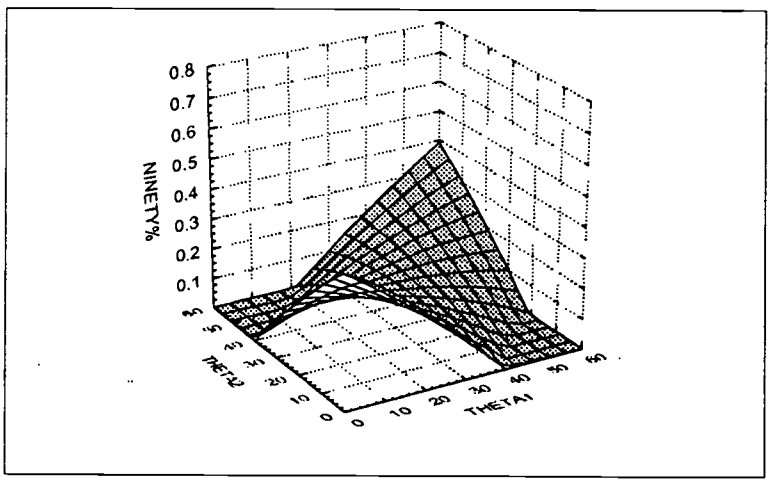
Figure 7
Overlap Rates for Davey-Parshall Exposure Control



Median of Overlap Dist

Overall: 0.0791

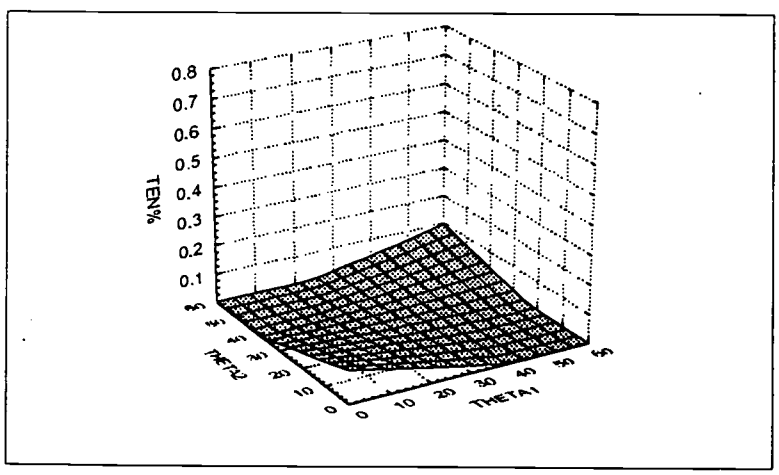
Similar: 0.1946
Different: 0.0646



90th % of Overlap Dist

Overall: 0.1647

Similar: 0.3076
Different: 0.1467



10th % of Overlap Dist

Overall: 0.0251

Similar:
Different: 0.0901
out: 0.0169

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

TM028851

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Test Development Exposure Control for Adaptive Testing

Author(s): Parshall, C G, Davey, T, & Nering, M

Corporate Source: NCME

Publication Date: April, 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: C G Parshall

Printed Name/Position/Title: C G Parshall
Psychometrician

Organization/Address: HMS 401, USF, Tampa, FL 33620

Telephone: 813/974-1256

FAX: 813/974-5132

E-Mail Address: parshall@seaweed.coedu.usf.edu

Date: 5-11-98

(over)